Mapping Spectral Frames to Pitch with the Support Vector Machine

Andrew W. Schmeder
Center for New Music and Audio Technology, University of California at Berkeley
andyschm@cnmat.berkeley.edu

Abstract

The Support Vector Machine algorithm is studied in the context of pitch estimation. Its learning capacity is analyzed using an artificial dataset of harmonic spectra. We propose an architecture for learning pitch in difficult real-world scenarios, and demonstrate its application with a database of guitar sounds. Domain-specific aspects of kernel methods are discussed, and a method for extracting structural knowledge via visualization is examined.

1 Introduction

Pitch estimation has a rich legacy of research spanning several decades, and yet remains a musically poignant subject. We see the latency of existing methods (Yoo and Fujinaga 1999) as a major obstacle in real-time performance (Wessel and Wright 2002), and therefore seek to minimize it to below perceptible levels. Motivation to apply the Support Vector Machine comes from the hope that, given its adaptive capacity to absorb knowledge in context and to produce sparse solutions compatible with real-time implementation, it may be possible to obtain estimations which are useful under aggressive latency constraints.

2 The Support Vector Machine

The Support Vector Machine (SVM) is currently a popular statistical learning algorithm based on the principle of structural risk minimization (Schoelkopf and Smola 2002). The SVM is a "kernelized" non-linear extension of the well known Perceptron algorithm (Cristianini and Shawe-Taylor 2000) and can be adapted to perform classification, regression or novelty detection.

The SVM is a maximum-margin classifier, and shows its strength in classification problems. However, for this paper we focus on ϵ-insensitive Support Vector Regression (ϵ-SVR), primarily in order to make a more clear analysis of performance. It is not the intention of this paper to provide a complete mathematical background; however some details are presented for clarity.

Given a function $f(x) : \mathcal{X} \to \mathbb{R}$ (in our case $\mathcal{X} = \mathbb{R}^d$), a sampling of its distribution, $\{(\mathbf{x}_i, y_i) : i = 1 \dots N\}$, and a kernel function $K(\mathbf{x}, \mathbf{y})$, ϵ-SVR finds $\mathbf{a}$ and b so that an approximation of f is constructed with minimal error.

$$f(\mathbf{x}) \approx \sum_i a_i K(\mathbf{x}, \mathbf{x}_i) + b : \mathbf{x}_i \text{ is a } \textit{support vector}$$

A commonly used kernel for real-valued data is the radial basis function (RBF).

$$K(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|^2}$$

The *kernel trick* is that $K(\mathbf{x}, \mathbf{y}) =< \Phi(\mathbf{x}), \Phi(\mathbf{y}) >$, where $\Phi(\mathbf{x}) : \mathcal{X} \to F$ for F a very high (possibly infinite) dimensional feature space with dot product $< \cdot, \cdot >$, i.e., the unique solution is reached via dual optimization under the *implicit* map Φ. Owing to the enormous dimensionality of F, a good linear model is easy to find; a realization of the blessing of dimensionality (Gershenfeld 1998).

2.1 Applications in Signal Processing

The SVM has recently been successfully applied in many audio processing tasks (Bousquet and Perez-Cruz 2003), including novelty detection (Davy and Godsill 2002), segmentation of silence, speech, music and noise (jiang Zhang, Lu, and Li 2001), and chirp classification (Gretton et al. 2001).

Effective use of similarity metrics with time-domain signals requires transformation into a translation-invariant domain, such as Cohen's class TFR / DFT, or customized wavelet-type bases (e.g., as in Krishnapuram and Carin (2002)). For pitch recognition, we are primarily interested in a collection of steady-state sinusoids, so the DFT is an apt representation. To enhance the sharpness of peaks, we also analyze incremental phase differentials to reassign energy towards the actual peak. This helps improve sparsity of the input vectors, which makes the SVM faster to solve.

2.2 Feature Scaling and Kernel Parameters

The proper choice of kernel parameters (e.g., γ, ϵ) is important, but is relatively easy. In difficult problems the issue of feature selection or feature scaling also arises, since unbalanced or unnecessary features may skew the results (Chapelle et al. 2002).

We found the standard rules-of-thumb (e.g., unit-variance normalization) to have a detrimental effect when applied to spectral frames, so we normalize to induce amplitude invariance but do not rescale. A Simplex optimizer was used to choose the best kernel parameters, and was also allowed to perform limited rescaling of the data without compromising tractability.

2.3 Spectral Frame SVM

Herein we will refer to our basic pitch estimation method as Spectral-Frame SVM (SF-SVM), denoting the systematic combination of normalized phase-enhanced power spectrum preprocessing with RBF kernel and Simplex wrapper optimization.

3 Learning Capacity Evaluation

This section analyzes the performance of the SF-SVM in pitch estimation over an artificial dataset under degrading conditions. For sake of comparison, we also present results with other methods: Maximum-Likelihood (ML) (de la Cuadra, Master, and Sapp 2001); Trained-ML (TML), an example-based version of the prior; and *fiddle˜* (Puckette, Apel, and Zicarelli 1998). All methods are real-time capable.

3.1 Experiment Design

100 time-domain harmonic signals are generated, each consisting of 15 partials over a fundamental ranging 40–52 (in terms of whole MIDI note numbers). Phase of each partial is random, and amplitude decays quadratically.

Pitch estimation errors are recorded for each example. For trained methods (TML and SF-SVM) error is taken from the unseen test sets in 3-fold cross validation.

3.2 Results

Error is measured in units of the 12-tone scale (1.0 = 1 semitone). Figures below show log of the Mean Squared Error (MSE), with bezier smoothing to enhance readability.

Window Size. Reduction in window size is correlated with loss of resolution in the frequency domain while improving the latency characteristics. We test the capacity to detect F_0 of noise-free harmonic spectra at window sizes ranging from 3–192 msec (FS = 44.1khz).

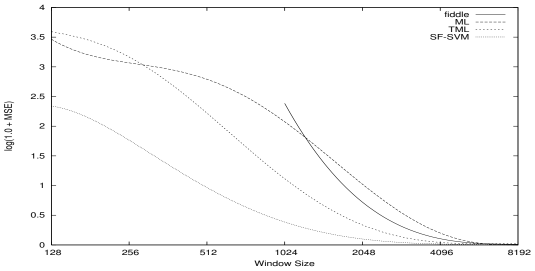

Figure 1: Window Size vs. log MSE.

Spectral Line Broadening. Real-world instruments exhibit non-linearities which may manifest in the form of spectral line broadening. We model this by modulating the frequency of each partial with noise. All methods use a 2048-sample window in this test.

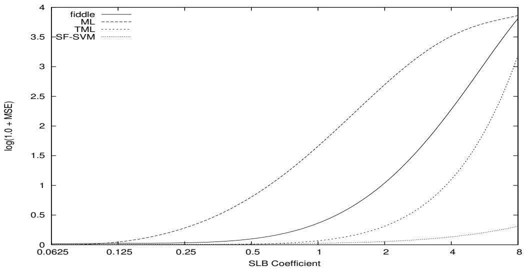

Figure 2: Spectral Line Broadening vs. log MSE.

Conclusion. These benchmarks should not be interpreted as normative, however the SF-SVM is a clear winner. This is not surprising considering it significantly more sophisticated than ML or TML and has the advantage over *fiddle˜* of example-based learning. It is especially interesting to note that SF-SVM delivers competitive results while cutting the latency requirement approximately in half.

4 The X40 Hexaphonic Guitar Dataset

Professional guitarist John Schott was employed to produce recordings suitable for testing our learning system. For each of six strings, every note (over 18 frets) was plucked forty times, cleanly separated by silence. The output was recorded via bridge-mounted piezo pickups in six channels.

A custom interface was created in Pure Data (Puckette 1996) to review each event, to prune mistakes and to match a pitch using a synthesized reference tone by ear. Onsets are located based on amplitude threshold, and time-domain frames are taken from immediately after this point, then transformed to normalized power spectra as described earlier and labeled to construct the final dataset. Herein we refer to this as the X40 dataset.

5 Toward Effective Pitch Classification

The perceptual cost of pitch misclassification in musical applications is high, therefore we need to build an estimator which is capable of note classification with nearly complete certainty.

Direct application of ϵ-SVR to the X40 dataset does not yield sufficient accuracy. Furthermore, training time grows quadratically or worse with respect to the number of examples, a problem which is magnified by use of iterative wrapper optimization.

Hierarchical Classification. One approach to training with large scale datasets is to use an ensemble of local models. This can be shown to produce results at least as good as training a single model with much improved training times and the added benefit of parallel computation (Collobert, Bengio, and Bengio 2002).

In the pitch classification problem, the classes are not completely independent; there exists macro-scale as well as local structure. Therefore we propose a reductionist multiple-model system, akin to a bisection search. We train a hierarchy of classifiers, each one providing a progressive refinement until the smallest scale is reached, at which point a single model can can decide between two neighboring pitches with high precision.

This method retains tractability, in both training and evaluation, for very large datasets. Error penalization in large-scale bisection decisions is relaxed by appropriate setting of the ϵ parameter, thereby retaining model sparsity. It is also interesting to note that the system benefits from training local models first. Because support vectors concentrate information, it is only necessary to retain the support vectors to train the models at the next larger scale [1], i.e., each level of training acts as a filter to reduce the size of the working set for the next level. Finally, the persistence of support vectors across multiple models means that some kernel function evaluations may be reused during model evaluation.

Experimental Worst-Case Bounds. We focus on the lowest string of the X40 dataset because it is the most difficult to learn and gives an upper bound on the performance of the overall system. In Figure 3 we clearly see that multi-model bisection classification is more than sufficient to achieve nearly 100% classification accuracy.

Range	ϵ	Bias	Std.	Confidence	# SV
40–41	0.5	0.1	0.07	99.99%	16
40–43	1.0	0.27	0.19	99.7%	29
40–47	2.0	0.32	0.33	99.99%	33
40–57	8.0	1.15	1.25	99.99%	54

Figure 3: Bisection classification confidence with 1024-sample window on low E string.

For higher pitch ranges it is trivial to obtain equivalent confidence with much lower latency; as low as 3 msec for the highest string (range 74–81).

6 The Geometry of Pitch-Space

Do kernel methods truly possess a capacity for insight? How is the SV method different from the maximum-likelihood template matching system? More generally, if the SVR is capable of extracting knowledge from the data, to what extent is that knowledge accessible to the researcher? To explore these questions, we examine the structure of the kernel matrix.

Kernel Principle Components Analysis. Kernel Principle Components Analysis (KPCA) (Mika et al. 1999) takes the kernel matrix $\mathbf{K}$, where $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, and performs classical PCA to yield up to N eigenvector/value pairs in F.

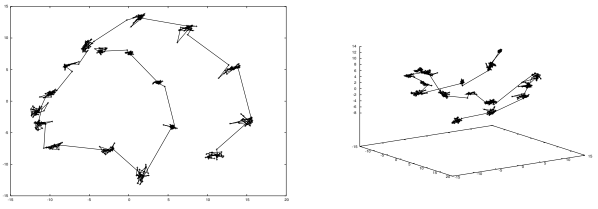

Figure 4: 3D embedding by projection onto primary eigenvectors of the kernel matrix, top and perspective view.

[1] This is not a rigorous statement, but is empirically grounded.

By projecting data from $\mathcal{X}$ onto the first few principle components in F, a dimensionality reduction is obtained. This operation was performed on the X40 dataset to produce an embedding in three dimensions. The result is an object familiar to pitch perception theory, a helix (Figure 4). This visualization gives an idea of how the weighted kernel projection invokes a hyper-linear model; the helix "linearizes" the spectral relationships, producing a consistent concept of pitch ordering.

7 Conclusion

We have demonstrated that the SVM can map spectral frames to pitch with high precision while requiring less than half the latency of competing methods. A bisection-search multi-model approach was proposed for use with large scale datasets, with experimental results. The ability to examine the geometry of data in feature space was demonstrated, producing a familiar structure reinforcing our intuition regarding pitch metrics.

8 Implementation

Programming for this project was done in Python using numerical and scientific computing modules on a Linux platform. LIBSVM (Chang and Lin 2001) was used as the basis for the SVR, and Gist (Noble and Pavlidis 2002) for Kernel PCA.

9 Acknowledgments

This project would not have been possible without the support and inspiration of David Wessel and the insight of Adrian Freed. This work was supported in part by Gibson Guitar.

References

Bousquet, O. and F. Perez-Cruz (2003). Kernel methods and their applications to signal processing. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Chang, C.-C. and C.-J. Lin (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapelle, O., V. Vapnik, O. Bousquet, and S. Mukherjee (2002). Choosing multiple parameters for support vector machines. *Machine Learning 46*(1-3), 131–159.

Collobert, R., Y. Bengio, and S. Bengio (2002). Scaling large learning problems with hard parallel mixtures. *Lecture Notes in Computer Science 2388*, 8–23.

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods).* Columbia University Press.

Davy, M. and S. Godsill (2002). Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE.

de la Cuadra, P., A. Master, and C. Sapp (2001). Efficient pitch detection techniques for interactive music. In *Proceedings of the International Computer Music Conference*. International Computer Music Association.

Gershenfeld, N. (1998). *The Nature of Mathematical Modeling*. Cambridge, MA: The MIT Press.

Gretton, A., M. Davy, A. Doucet, and P. J. W. Rayner (2001). Nonstationary signal classification using support vector machines. In *11th IEEE Workshop on Statistical Signal Processing*, pp. 305–308. IEEE Signal Processing Society.

jiang Zhang, H., L. Lu, and S. Z. Li (2001, May 20). Content-based audio segmentation using support vector machines.

Krishnapuram, B. and L. Carin (2002). Support vector machines for improved multiaspect target recognition using the fisher kernel scores of hidden markov models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Mika, S., B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch (1999). Kernel PCA and de–noising in feature spaces. In M. S. Kearns, S. A. Solla, and D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11*. MIT Press.

Noble, W. S. and P. Pavlidis (2002). Gist: Support vector machine and kernel principal components analysis software toolkit. Software available at `http://microarray.cpmc.columbia.edu/gist`.

Puckette, M. (1996). Pure data: another integrated computer music environment.

Puckette, M., T. Apel, and D. Zicarelli (1998). Real-time audio analysis tools for pd and msp. In *Proceedings of the International Computer Music Conference*. International Computer Music Association.

Schoelkopf, B. and A. J. Smola (2002). *Learning with Kernels*. Cambridge, MA: The MIT Press.

Wessel, D. and M. Wright (2002). Problems and prospects for intimate musical control of computers. *Computer Music Journal 26*(3), 11–22.

Yoo, L. and I. Fujinaga (1999). A comparative latency study of hardware and software pitch-trackers. In *Proceedings of the International Computer Music Conference*. International Computer Music Association.